Journal of the Text Encoding Initiative
Rolling Issue | 2019

The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in TEI XML

Christof Schöch, José Calvo Tello, Ulrike Henny-Krahmer and Stefanie Popp

Electronic version
URL: http://journals.openedition.org/jtei/2085
DOI: 10.4000/jtei.2085
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Christof Schöch, José Calvo Tello, Ulrike Henny-Krahmer and Stefanie Popp, « The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in TEI XML », *Journal of the Text Encoding Initiative* [Online], Rolling Issue, Online since 14 August 2019, connection on 14 March 2020. URL : http://journals.openedition.org/jtei/2085 ; DOI : https://doi.org/10.4000/jtei.2085

The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in TEI XML

Christof Schöch, José Calvo Tello, Ulrike Henny-Krahmer, and Stefanie Popp

ABSTRACT

The CLiGS textbox is published by the Computational Literary Genre Stylistics (CLiGS) group. The textbox is the group's publication channel for several collections of literary texts. We describe the rationale for the manner in which the collections of literary texts included in the textbox have been compiled, annotated, and published. Furthermore, we suggest several ways in which the text collections can be used for research in literary studies. We aim to document some of the work of the CLiGS group, to showcase the unique TEI XML-based collections of French, Spanish, Spanish-American, and Portuguese novels and French drama we make available, and to encourage reuse of

these text collections by others. We argue that agreement on common formats and procedures for text preparation, encoding, and publication fosters the accessibility, analysis, and reuse potential of literary text collections.

INDEX

Keywords: text collections, text curation, Spanish literature, French literature, Portuguese literature, metadata, usage scenarios

ACKNOWLEDGEMENTS

This research was funded by the German Federal Ministry of Education and Research (BMBF) under funding identifier 01UG1508.

1. Introduction

1 The CLiGS textbox is dedicated to making collections of literary texts in romance languages freely available. CLiGS (Computergestützte literarische Gattungsstilistik / Computational Literary Genre Stylistics)[1] is an early-career research group funded by the German Federal Ministry for Research and Education (BMBF). In this interdisciplinary group, four doctoral and three postdoctoral researchers with backgrounds in literary studies, computational linguistics, and text mining investigate the relationship between style and genre in literary texts, working primarily with collections of French and Spanish novels and plays. Currently, nine collections are available, consisting of Spanish novels, Spanish-American novels, French novels, French novellas, French drama, and Portuguese novels, respectively. The texts are published in TEI XML and plain text versions and include detailed document-level metadata.

2 The aims of the textbox are at least twofold: first, it aims to provide researchers in Romance studies with medium-sized, carefully encoded collections of literary texts and detailed metadata to enable them to do further research like the evaluation of authorship attribution methods or contrastive analysis of literary subgenres. Second, the textbox gives insight into the ways we work with text collections and metadata in our research group. By publishing our data and metadata, we hope

to receive feedback on our practices and to inspire other groups to share their data openly to the benefit of all. We also aim to foster digital research into the areas that the textbox covers, in particular French and Spanish literature.

3 Note that this paper refers to release 4.0.0 of the textbox, published in May 2018.[2]

2. Text Selection and Preparation

4 This section provides an overview of the principles of text selection that guided the creation of the textbox collections, describes the subset of TEI elements used to encode the texts, and summarizes some of the quality checks the texts have undergone.

2.1 Principles of Text Selection

5 The textbox currently contains novels, novellas, and short stories published between 1830 and 1940 in France, Italy, Spain, Portugal, and Spanish-America, as well as plays published between 1640 and 1680 in France, with a total of 388 texts or about 13.6 million words. With various usage scenarios in mind, each collection has been compiled in a slightly different manner. Note that this compilation of texts deliberately aims to provide medium-sized collections balanced around certain criteria, so the collections are not representative samples of the overall production of novels, novellas, short stories, or plays in the respective language and period. The text collections published in the textbox are subsets of larger collections being prepared by the group in the context of doctoral and postdoctoral research. It was decided to publish smaller subsets of the overall collections at an early stage, as a proof of concept, in order to receive feedback early on and to allow reuse by others even before the preparation of the full collections is concluded.

6 The two collections of Spanish novels currently available, the Corpus of Spanish Novels from 1880–1940 and the Collection of 19th Century Spanish-American Novels (1880–1916), have been prepared to be used for authorship attribution in Spanish. Accordingly, the two collections have been balanced with regard to the number of texts from different authors. They contain 24 Spanish novels published from 1880 to 1940, with three novels each from eight different authors and a total of 1.8 million words. Similarly, there are 24 Spanish-American novels published from 1880 to 1916, with three texts per author and a total of 1.1 million words.

7 The Corpus of Spanish Short Stories from 1880–1940 contains 20 texts written by 8 Spanish authors (Bazan, Blasco Ibáñez, Clarín, Galdós, Miró, Pereda, Unamuno y Valle). In total this collection contains 302 short stories, which represent 811,000 tokens. It was originally a subset of a greater corpus of Spanish prose that was finally split into two corpora (novels and short stories). One possible use is to analyze the style of the same author in novels compared to short stories.

8 The Collection de romans français du dix-neuvième siècle has been compiled to enable contrastive analyses regarding major subgenres of the novels across several decades. With this in mind, 36 French novels first published in the 1860s, 1870s, or 1880s and belonging to the subgenres of adventure novel, crime fiction, *Bildungsroman*, and fantastic novel have been selected, with each genre and decade covered by similar amounts of text. This collection contains a total of 4.3 million words.

9 In a similar manner, the Collection de nouvelles françaises du dix-neuvième siècle contains 28 texts published in the 1830s, 1840s, 1850s, 1860s, 1870s, 1880s, and 1890s that can be classified as either fantastic or realistic novellas. With its wider chronological span, this collection is suitable not only for contrastive analyses of fantastic and realistic types of novellas, but also for investigations into the diachronic development of the genre. As novellas are much shorter than novels, this collection contains a total of slightly less than 500,000 words.

10 The Collection de pièces de théâtre français du dix-septième siècle contains 100 dramatic works first performed between 1640 and 1670 and classified as being either comedies, tragedies, or tragicomedies. For better comparability, all plays selected are written in verse. This collection is a subset of the Théâtre classique collection edited by Paul Fièvre (2007–2018) and is suitable for contrastive analyses of dramatic subgenres, particularly for investigations into the particular position of tragicomedy with regard to comedy and tragedy. This collection contains about 1.3 million words.

11 The textbox contains two collections of Italian texts: the Collection of Italian Short Stories and Novellas (1880s–1920s) and the Collection of Italian Novels. The collection of short stories contains 90 texts written by three authors. It amounts to approximately 300,000 words and can, for example, be used to test authorship attribution. The corpus of Italian novels includes 21 texts by 15 authors, written between 1850 and 1915. It contains about 2.3 million words.

12 The Collection of 19th Century Portuguese Novels (1840–1910) contains 30 novels written by 14 male Portuguese authors and published between 1840 and 1910. It was initially created to support comparative research for historical versus non-historical novels, but has been extended for use with topic modeling. As a consequence, the collection contains many historical novels but is on the whole not balanced with regards to authors, decades, or subgenres. Hence, for each individual reuse, an appropriate subselection or adaption of the set of novels may be necessary. The collection contains a total of 2.5 million words.

13 The main sources of the texts included in the textbox at this point are web portals and similar pages which hold HTML, EPUB, or other full-text versions of the novels, novellas, and short stories. Additionally, some sources which offer texts as PDF files or images have been used. Sources for French, Spanish, Italian, and Portuguese texts are Wikisource and Project Gutenberg. Among the sources for the French texts are Ebooks libres et gratuits, La Bibliothèque électronique du Québec, ÉFÉLÉ, and Théâtre classique. Spanish and Spanish-American novels and short stories have been obtained from the Biblioteca Virtual Miguel de Cervantes, Biblioteca Digital Hispánica, and Biblioteca Digital Argentina, among others. An important source for the Italian novels was Liber Liber. The Portuguese novels come from Luso Livros. (See Appendix 1 for a non-exhaustive list of sources for literary texts in Romance languages.)

2.2 Use of TEI XML

14 All incoming texts are prepared according to a common data schema established by the CLiGS group. Most of the texts are made available by their sources as plain text, HTML, or EPUB, and need to be transformed to TEI XML, which we do with Python scripts (using a combination of XPath and regular expressions) tailored to the different sources and the different formats. In the case of the source files from Théâtre classique, the original TEI P4 files have been transformed to TEI P5 using an XSLT-based transformation. The same schema is shared across all collections and is published in a separate CLiGS repository.[3] This ensures that any processing or analysis script designed with one collection in mind can easily be reused when working with another collection.

15 The base format builds on XML mainly for two reasons. First, XML allows for a direct integration of metadata and textual content and not only helps to ensure the integrity of both, but also allows for structured comments and qualifications on individual metadata items (indicating, for example,

the person responsible for entering a value or the degree of certainty of a value), which is not as easily achieved in tabular formats. Second, XML provides us with convenient mechanisms to preserve structural information contained in the source files (like divisions into front, body, and back matter, chapters and paragraphs, or acts and scenes) or inline typographical information (like italics or bold type), which might be of interest for text analysis.

16 The data schema for the textbox follows the TEI Guidelines (for an introduction, see Burnard 2014), thus connecting the resources of the textbox to an established infrastructure and *de facto* standard as well as to a large community of users. The schema includes elements and attributes from the TEI modules `core`, `header`, `textstructure`, `analysis`, `drama`, `namesdates`, and `linking`, being as restrictive as possible. It was decided to include basic elements for the encoding of literary texts in prose and verse-like paragraphs and verse lines, but to avoid more specialized block-level elements like lists and tables or inline elements like `<foreign>` or `<emph>` in order to keep things simple.[4]

17 All the elements of the CLiGS encoding scheme conform to TEI version P5. A few attributes have been added to the schema, though, using a project-specific namespace. Partly, these are attributes that result from the XML output of linguistic annotation and that could not be mapped to TEI. In addition, a new attribute is used to allow for the expression of degrees of importance to assignments of metadata categories (see section 3.2 below for details).

18 When compared to other TEI customizations, it can be stated that the CLiGS schema is closely related to the DTA-Basisformat (DTABf). The DTABf is a subset of the TEI that serves as the basis for the annotation of full texts in the German Text Archive (see Haaf, Geyken, and Wiegand 2014–2015). Disregarding minor differences in the definition of allowed attributes and attribute values, the CLiGS schema can be considered a subset of the DTABf. As to TEI Lite or TEI Simple (see Burnard and Sperberg-McQueen 2012; Mueller et al. 2015), the CLiGS encoding schema is not an exact subset of either of them. In comparison to the CLiGS schema, these established TEI customizations include a larger range of elements and attributes and would have allowed for a greater degree of variation in the encoding. On the other hand, these customizations would not have covered the case of embedded texts, a recurring phenomenon in the narrative texts we publish, where, for example, letters, newspaper articles, and historical documents interrupt the flow of the main text. To cover these cases, the element `<floatingText>` has been included in the CLiGS schema.

19 Apart from the reference versions of the texts in TEI XML with structural annotation and embedded document-level metadata, we also provide derived versions for various usage scenarios. We provide some sample application scenarios using these different formats in section 5 below.

20 In addition to the reference format, each collection is made available in a simple plain text format automatically derived from the TEI version, containing only the text included in the body of the narrative texts and plays (in particular, excluding prefaces and other paratext as well as notes) and with external metadata provided in tabular format. This format is especially suitable for direct use with the stylo package for R (Eder, Kestemont, and Rybicki 2016) and other tools operating on the surface level of texts.

21 Moreover, the collections of French, Spanish, Spanish-American, Italian, and Portuguese novels, novellas, and short stories are made available in a version combining basic structural markup (chapter and sentence divisions) with token-level linguistic annotation (including lemma, part-of-speech, morphology, and basic semantic annotation using FreeLing and WordNet). We decided to use the tagger of the NLP package FreeLing (see Padró and Stanislovsky 2012) for the linguistic annotations because its tagset is quite fine-grained and it comprises WordNet-based sense annotation and disambiguation (on WordNet in general, see Miller 1995 and Fellbaum 1998). We created a workflow to integrate the results of the annotation process into the TEI files. The result is a TEI file with the same header as the reference file, but with a different text body. Chapter divisions are preserved to allow for chapter-level analyses. Inside each division, the text is broken up into sentences and words, carrying the results of the linguistic annotation process as attributes. The attributes that could be mapped to TEI were kept in the TEI namespace. The remaining attributes were encoded in the CLiGS namespace (see example 1).

Example 1. A sentence in the linguistically annotated TEI derivative (“Pasaron dos días.”).

```
  <s xmlns:cligs="https://cligs.hypotheses.org/ns/cligs">
   <w cligs:form="Pasaron" lemma="pasar" cligs:tag="VMIS3P0" cligs:ctag="VMI"
pos="verb" type="main" cligs:mood="indicative" cligs:tense="past" cligs:person="3"
cligs:num="plural" cligs:wnsyn="00339934-v" cligs:wnlex="verb.change">Pasaron</w>
   <w cligs:form="dos_días" lemma="TM_d:2" cligs:tag="Zu" cligs:ctag="Zu"
pos="number" type="unit" cligs:wnsyn="xxx" cligs:wnlex="xxx">dos_días</w>
   <w cligs:form="." lemma="." cligs:tag="Fp" cligs:ctag="Fp" pos="punctuation"
type="period" cligs:wnsyn="xxx" cligs:wnlex="xxx">.</w>
  </s>
```

22 The linguistically annotated format is suitable for direct import into the TXM desktop environment—a tool for text analysis that is capable of performing complex queries on such annotations (see Heiden 2010 and the website of the Textométrie project[5]).

23 Finally, the collection of French plays is available not only in TEI, but also in the "Zwischenformat" developed by the DLINA group.[6] This format represents an abstraction from the full TEI XML in that it maintains the plays' structural division into acts and scenes but replaces the speaker text with statistics regarding the number and length of the speech acts of each speaker in each scene, which notably makes possible the efficient calculation of network characteristics of a play.

2.3 Quality Control

24 Joining many texts from various sources into one collection may lead to a group of texts that is heterogeneous in vocabulary, spelling, and text quality in general. It also entails the risk of carrying over errors from the sources that may remain undetected but might influence the results of text analyses. General kinds of mistakes like structural and orthographic errors may be introduced by an OCR process. But there might also be other, more source-specific kinds of errors. The texts are checked for completeness (so that, for example, they contain all the chapters they should) and for conformance of the TEI encoding to the custom TEI schema. Additionally, as a simple way to check the quality of the texts that go into the textbox on the orthographic and the character level, a dedicated spellcheck routine was implemented in Python using the "pyenchant" package (see Henny and Schöch 2016).

25 To account for named entities, foreign words, and other special cases that are not covered by the standard spellchecker but should count as legitimate words in the texts, the spellchecking script was combined with several lists of exception words. The remaining errors are counted for each text and for the collection as a whole. Subsequently, they are stored in an error list. Such a list may just provide information about the reliability of the texts in terms of errors on and below the word level, or it might be the basis for correcting frequently recurring errors. To give an example, there were around 8,000 different errors in the collection of French nineteenth-century novels when the spellcheck was applied for the first time. After taking into account named entities, foreign words,

and some dialectal and colloquial words, most of the remaining errors did not occur more than once. The quality of the texts was better than initially thought and fixing the few recurring errors was feasible.

3. Types and Implementation of Metadata

26 In this section we provide an overview of the kinds of metadata provided for the text collections, explain how the metadata is implemented in the TEI header section, and argue for the usefulness of this approach for further processing of the texts. When the types of metadata are presented in section 3.1, XPath expressions are given to illustrate how they are implemented, while the overall implementation strategy is explained in section 3.2.

3.1 Types of Metadata

27 Following the classification of the NISO (2004), the collections provide two different kinds of metadata: descriptive and administrative. The descriptive metadata document information about four main areas:

1. *Authorship:* A reference to the author is kept in three different ways: as the full name (`//titleStmt/author/name[@type='full']`), as a short reference unique in the collection to identify the author in analyses (`//titleStmt/author/name[@type='short']`), and using the VIAF identifier (Virtual International Authority File,[7] `//titleStmt/author/idno[@type='viaf']`) to allow unambiguous reference beyond our project, make it easier for other projects to reuse this corpus, and allow the retrieval of additional information about the authors. Together with this information, we document the continent and country (`//textClass/keywords/term[@type='author.continent']` and `//textClass/keywords/term[@type='author.country']`, respectively) to which each author is related (in most if not all cases, this is the country where they were born) and their gender (`//textClass/keywords/term[@type='author.gender']`).

2. *Text*: Similarly to the metadata about the author, the full title (`//titleStmt/title[@type='main']`), a short reference unique to the collection (`//titleStmt/title[@type='short']`), and the VIAF identifier or the identifier from the relevant National Library in case the work does not have an identifier in VIAF (`//titleStmt/title[@type='idno']/idno[@type='viaf']`) are recorded for the text. This identifier is a reference to the work (not a specific expression or manifestation of it) which makes it possible to connect this work instance to other instances outside of the project. Metadata about the text also indicate its extent according to various measures like words (`//extent/measure[@unit='words']`) or characters (`//extent/measure[@unit='chars']`) for use in the analysis, as well as the optional secondary title of the work, which often contains a genre label (`//titleStmt/title[@type='sub']`). The source of the text is referenced in three different ways in order to differentiate and control the various steps in the process of establishing the text: a reference to the digital source forming the basis for our digital version of the text (`//sourceDesc/bibl[@type='digital-source']`); if available, a reference to the print edition that the digital source used (`//sourceDesc/bibl[@type='print-source']`); a reference to the first publication of the text as a monograph or in a journal (`//sourceDesc/bibl[@type='first-edition']`), especially useful for assigning a date of publication to the text. In cases where no digital source was available, the CLiGS version of the text is based directly on a print source. In some cases, the print source is itself the first edition of the work. Because of the different possible constellations of source texts and to ensure comparability across the collection, all three bibliographic descriptions of source texts are provided in every file.

3. *Genre*: Since the main focus of the CLiGS project is literary genre, a considerable part of the metadata is directly connected to it. Any reference to a genre in the title of the work is formally collected as either the *title genre* (`//textClass/keywords/term[@type='text.genre.title']`) or the *title subgenre* (`//textClass/keywords/term[@type='text.genre.subgenre.title']`). Besides that, we have established a hierarchical system of *supergenre* (in which the great majority of our texts are designated as "narrative," the remainder as "drama": `//textClass/keywords/term[@type='text.genre.supergenre']`), *genre* (that is, novels or novellas: `//textClass/keywords/term[@type='text.genre']`), and *subgenre* (the subtype of the novel, for example adventure novel or political novel, or the subtype of the drama, for example comedy and tragedy: `//textClass/keywords/term[@type='text.subgenre']`). Since these collections have been established to study the novel's subgenres, the most important level of this hierarchy is the subgenre. Here, it is possible to assign multiple different values to a given novel in order to account for cases where novels can usefully be described as hybrids of several different subgenres. These multiple values are also structured in a way that requires one subgenre value to be designated as the main subgenre, with the other values designating secondary subgenres. This structure represents a more nuanced and realistic representation of the relations between works and genres, but it also allows working with a single subgenre concept per text (see section 3.2 for implementation details). Together with this information, we also provide some additional descriptive information like form (prose or verse: `//textClass/keywords/term[@type='text.form']`) and the primary publication format (normally, as a monograph: `//textClass/keywords/term[@type='text.publication.type']`).

4. *Content of the text*: Finally, regarding the Spanish and Spanish-American novel collections, we collect different metadata about the content and meaning of the text: an optional summary to give an overview of the text (`//profileDesc/abstract`); the narrative perspective, an aspect that has a great impact on the frequency of the pronouns and verbs used in the text (`//textClass/keywords/term[@type='text.narration.narrator']`); the gender of the protagonist (`//textClass/keywords/term[@type='text.characters.protagonist.gender']`); the kind of place where the novel's action takes place primarily (city or rural; `//textClass/keywords/term[@type='text.setting.settlement.type']`). These last two metadata items are part of many definitions of subgenres of the novel, so it could be particularly useful for this information to be explicitly available.

28 Beyond descriptive metadata, the `<teiHeader>` also contains administrative metadata that help manage the collection and document the internal process of the creation of the documents.

1. The name of the editor of the TEI document (identified with an identifier that is used in other places in the document where the editor would like to make his or her responsibility for some information explicit (`//titleStmt/principal`), the legal status of the text (`//publicationStmt/availability`), a log of major changes, and the date when the document was created (both in `//revisionDesc`).[8]
2. Together with this information, an essential metadata item is documented: the text identifier (`//publicationStmt/idno[@type='cligs']`). These identifiers are built from two letters that summarize the name of the collection (for example `"rd"` for Romans français du dix-neuvième siècle, or `"ne"` for Novela española) and a number. The TEI file names are provided only with this identifier and the file extension corresponding to the format. Although this makes it harder for a human user to know which text each identifier refers to, we have found it extremely useful to have a simple way of identifying the text and of using its features and metadata. This allows us to write scripts that select, copy, or modify the files and prepare subcorpora made from a selection of texts in a collection and to use specific parts of the texts for particular experiments. Automatically renaming the files using a specific set of metadata is also possible, of course.

3.2 Implementation: The TEI Header with Keywords

29 With the TEI Header, the Text Encoding Initiative provides a sophisticated mechanism for recording metadata of textual resources (see Burnard 2014). The texts in the CLiGS collections use many of the TEI header's standard elements and attributes to record the information described in the previous section, especially the administrative metadata and the general descriptive metadata about the author and the work in the title statement and the source description. In the CLiGS project, further metadata are collected as a basis for the main application scenario: the classification of texts according to various factors such as author gender, author nationality, genre, subgenre, narrative perspective, and gender of the protagonist. The same kind of information is used to evaluate the results of text clusters and networks derived from textual similarities. From this perspective, specific metadata (about the author, genre, narrative strategies, and text content) contribute to the text classification. For example, if a text is written by an Argentine author, we can expect it to have different linguistic characteristics than a text written by an author from Spain, while a text with a first-person narrator will show a different usage of personal pronouns than a text narrated in the third person.

30 The TEI offers dedicated elements for some of these aspects, but not all of them. For the text collections at hand, it was important to keep the classification-related metadata in one place in order to facilitate queries for text analysis which can access the metadata item(s) that are relevant to the research question (e.g., authorship attribution, detection of author nationality, or genre classification). While this approach does not correspond to the common strategy followed, for example, in scholarly digital editions, where the focus is on the representation of the text, we believe that it is appropriate in a digital text collection created primarily for the purpose of text analysis. We decided to use the `<textClass>` element contained in the profile description to hold the classification-related metadata. Inside `<textClass>`, the `<keywords>` element is used, which, according to the TEI Guidelines, "contains a list of keywords or phrases identifying the topic or nature of a text."[9] In this case, the identification of topics is not the primary concern. Instead, the nature of the text is described by a set of controlled keywords. Each keyword is contained in a `<term>` element and the type of keyword is specified further in the `@type` attribute. The types of keywords are organized hierarchically, which is reflected in the structure of the attribute value: the different levels are separated by a dot. The main levels are author- vs. text-related keywords,

followed by sublevels (for example, `text.genre` and `text.narration`), a second layer of sublevels (for example, `text.genre.subgenre` and `text.narration.narrator`), and so on. The value of the keyword is given as the content of the `<term>` element. Example 2 shows the encoding of metadata from the Collection of 19th Century Spanish-American Novels (1880–1916):

Example 2. Detailed descriptive metadata using the `<keywords>` element. The example describes *La novela de la sangre* (1903) by Carlos Octavio Bunge.

```
  <textClass xmlns:cligs="https://cligs.hypotheses.org/ns/cligs">
   <keywords scheme="../keywords/keywords.xml">
    <term type="author.continent">America</term>
    <term type="author.country">Argentina</term>
    <term type="author.gender">male</term>
    <term type="text.language">Spanish</term>
    <term type="text.form">prose</term>
    <term type="text.genre.supergenre">narrative</term>
    <term type="text.genre">novel</term>
    <term type="text.genre.subgenre" cligs:importance="2" resp="#uhk"
cert="medium">historical</term>
    <term type="text.genre.subgenre" cligs:importance="1" resp="#uhk"
cert="medium">sentimental</term>
    <term type="text.narration.narrator">heterodiegetic</term>
   </keywords>
  </textClass>
```

31 Here, the first three elements give basic information about the author of the text, in this case the continent and country of birth or primary activity and the author's gender. Basic metadata about the author is important when texts from different continents and countries or from female and male authors are compared. To facilitate such analyses, basic metadata about the author is included in the keywords section of each text. The remaining terms are related to the text itself. The main language and form of the text are indicated, as well as information about the narrative perspective. In the example above, there are four terms referring to the genre of the text. In this case, the supergenre is `"narrative"` and the genre `"novel"`.

32 The subgenre is not limited to a single value. Instead, two assignments are made: `"historical"` and `"sentimental"`. Within the attribute `@cligs:importance`, numbers are given to express the importance of the assignment, a higher number meaning that an assignment is relatively more

important. This attribute has three different possible values: `"1"`, `"2"`, or `"3"`. If the text belongs to a single subgenre, the subtype value is `"3"`. If the text belongs to different subgenres, as in this example, the value of the subtype can only be `"2"` or `"1"`. For these cases, only one subgenre term may have a subtype with a value of `"2"`, and all the others need to be `"1"`. If there are only equally ranking subgenre assignments, they all have the value `"1"`. With this system, it is possible to describe texts as a mixture of genres that can be ranked. In the example, the novel is primarily a historical novel, but can also be considered a sentimental novel. The `@resp` attribute serves to indicate who is responsible for the subgenre assignment and the `@cert` attribute is used to express how confident the editor is in the information provided.

33 The types of keywords as well as their values are controlled in a taxonomy stored in a separate TEI file (`keywords.xml`), linked to from the `@scheme` attribute. The taxonomy is published together with each collection. The type values of the terms (e.g., `"genre.subgenre"`) correspond to the identifiers of the categories in the taxonomy. The hierarchy of term types indicated in the `@type` attribute corresponds to the hierarchical organization of categories in the external taxonomy. An excerpt from a keywords file is given in example 3:

Example 3. Some of the information included in the `keywords.xml` file.

```
<category xml:id="text.genre">
 <catDesc>text.genre</catDesc>
 <category xml:id="text.genre_1">
  <catDesc>novel</catDesc>
 </category>
 <category xml:id="text.genre_2">
  <catDesc>novella</catDesc>
 </category>
 <category xml:id="text.genre.subgenre">
  <catDesc>text.genre.subgenre</catDesc>
  <category xml:id="text.genre.subgenre_1">
   <catDesc>historical</catDesc>
  </category>
  <category xml:id="text.genre.subgenre_2">
   <catDesc>sentimental</catDesc>
  </category>
 </category>
</category>
```

34 The conformance of the metadata terms and values in the text files to the keywords taxonomy is checked with a Schematron file (`keywords.sch`) which is also published together with each collection. Example 4 shows how the subgenre values are controlled by the Schematron file:

Example 4. Excerpt from the `keywords.sch` file.

```
  <sch:rule context="//tei:term[@type='text.genre.subgenre']">
   <sch:assert test=". = document('keywords.xml')//tei:category
[@xml:id='text.genre.subgenre']/tei:category/tei:catDesc">
   Metadata error: text.genre.subgenre
   </sch:assert>
  </sch:rule>
```

35 We decided to use this system of correspondence to connect the keywords to the taxonomy instead of, for example, using `<catRef>` with the `@target` attribute because there are not only closed lists of values but also open lists of keywords. The subgenre labels derived from text titles, for instance, can be very diverse and are not controlled. With `<catRef>`, we would have been obliged to anticipate all the possible values for these categories in the external taxonomy. With the keywords system, the values can either be controlled or be left as an open list. A second argument for the keywords system is that it allows us to keep the encoding uniformly structured. With the `<catRef>` approach, it would have been necessary to find an alternative encoding for open lists of keywords, which would have complicated queries operating on the classification metadata. Third, it was important to us to keep all the metadata explicitly in every TEI file to facilitate reuse of the text collections. The corpora in the textbox are not conceived as closed databases with several interdependent components. The taxonomy of keywords and the schema controlling the keywords are auxiliaries to maintain the consistency of the classification metadata, but they are not necessary to understand and use individual TEI files in the collection. As the application scenarios are diverse, we want to enable users to freely select and recombine subsets of the collections without the need to keep track of external files. With the `<catRef>` approach, it would have been obligatory to take into account external category identifiers and descriptions in order to reuse the texts. The strategy described here for work-level metadata ensures that relatively detailed metadata can be stored in a consistent, well-documented way that is compliant with TEI XML.

4. Publication Strategy

36 The publication strategy for the CLiGS textbox collections relies on two infrastructures which, together, provide us with the flexibility we need and the guarantees for sustainable long-term archiving and access one can rightfully expect.

37 The text collections are curated and published using a public GitHub repository. Such a repository provides important features for publishing and archiving, such as fully automatic versioning (so that any changes to the texts, their markup, and the metadata are automatically documented), issue tracking (so that problems and possible enhancements of the collections can be suggested by any interested party as well as managed and resolved collaboratively), collaborative curation (so that all team members can contribute to building the text collections), and convenient sharing of collections, either by cloning or forking of Github repositories (enabling access to all versions and branches of a repository), or simply by downloading a ZIP archive of the most recent state of the repository on a one-time basis.

38 In addition, because GitHub as a commercial entity does not provide any guarantee for the long-term availability of the data, stable versions of the text collections are designated as releases (using semantic versioning) and archived on Zenodo.org, a long-term data and publications archiving service for researchers across Europe managed by OpenAire and supported by CERN (see Nielsen 2013). The textbox repository on Github is connected to Zenodo.org so that any new release is automatically archived and published on Zenodo.org. This includes the requirement that each release receives a DOI (Digital Object Identifier), providing the unambiguous identification of the resource, and ensures that the text collections will remain available for download even after the CLiGS project is terminated—or after GitHub ceases to exist. With this scenario in mind, it is important to include the DOI of the data in addition to the GitHub link whenever referencing a text collection. An additional benefit of the publication at Zenodo.org is that search engines, whether general-purpose or more specialized ones that focus on open-access scholarly resources and publications (for instance, BASE, the Bielefeld Academic Search Engine[10]), index Zenodo.org and thereby improve discoverability.

39 Another aspect of publication is licensing. All texts included in the CLiGS textbox are in the public domain, so we publish them without any licence-based restrictions and do not wish to impose any limitation on their use by others. The TEI markup, including the metadata, is published with

a Creative Commons Attribution license (CC-BY) solely in order to encourage users of the text collections to attribute academic credit to the group for the substantial editorial work which resulted in the TEI encoding and rich metadata. For the same reason and for convenience, we provide a citation suggestion for each collection.[11]

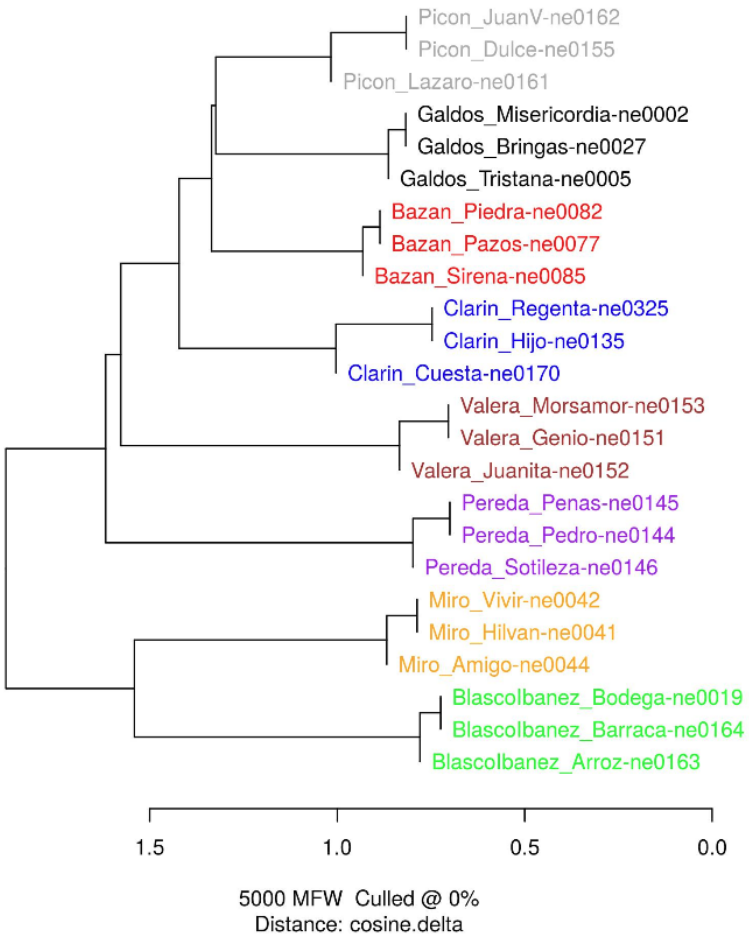
5. Usage Scenarios

40 The following section provides an overview of some of the usage scenarios which we aim to support with the textbox collections. These scenarios are intended not only as examples of analyses conducted by members of the CLiGS group, but also as suggestions for potential users of the CLiGS collections of texts. Authorship attribution, network analysis, textometric analysis, and topic modeling were chosen because they represent typical cases of quantitative text analysis in literary studies, the working area of the CLiGS group. Although the collections were primarily created to enable these types of analyses, the TEI files can of course be used in other scenarios as well.

5.1 Authorship Attribution

41 Among the many possible methods that can be applied to such collections are stylometric analyses for authorship attribution. For this purpose, we use the stylo package for the R statistical environment (Eder, Kestemont, and Rybicki 2016). Here, we used a custom implementation of the Cosine Delta Distance, proposed by Smith and Aldridge (2011) and discussed and tested by Evert et al. (2017). Distances were calculated based on the 5,000 most frequent words as features using the labeled plain text of the Corpus of Spanish Novels from 1880–1940 of the textbox:

Figure 1. Results of using Cosine Delta on the corpus of Spanish novels.



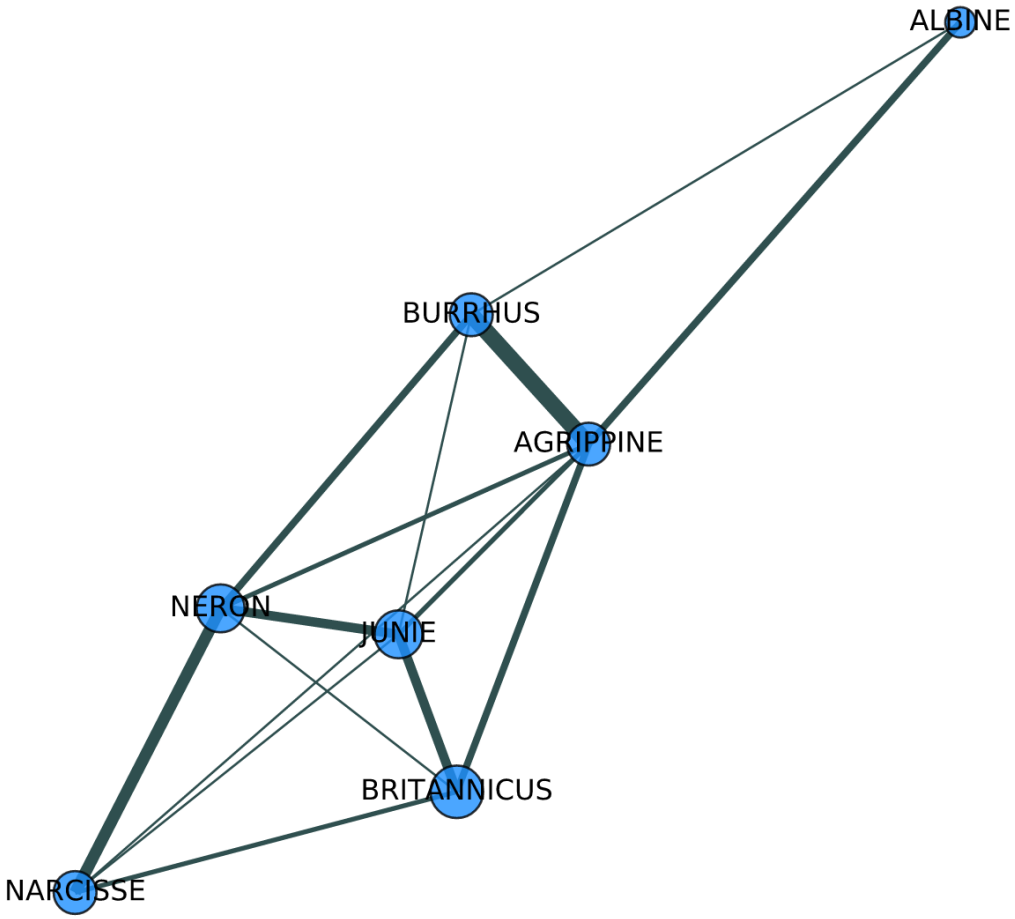
42 Using this version of Delta and this type and number of features, each set of three texts per author is clustered together, which indicates that the authorship is correctly recognized. This is not the case when applying other versions of Delta or other distance measures, or when significantly changing the number of features (Calvo Tello 2016). We invite other researchers to use these collections for further analyses.

5.2 Network Analysis

43 Another usage scenario supported by some of the text collections in the textbox concerns network analysis. The collection of seventeenth-century French plays is particularly relevant here, as the TEI markup makes the structure of the text (acts and scenes) as well as the interactions between speakers (who speaks how many lines and words in which scene) explicit.

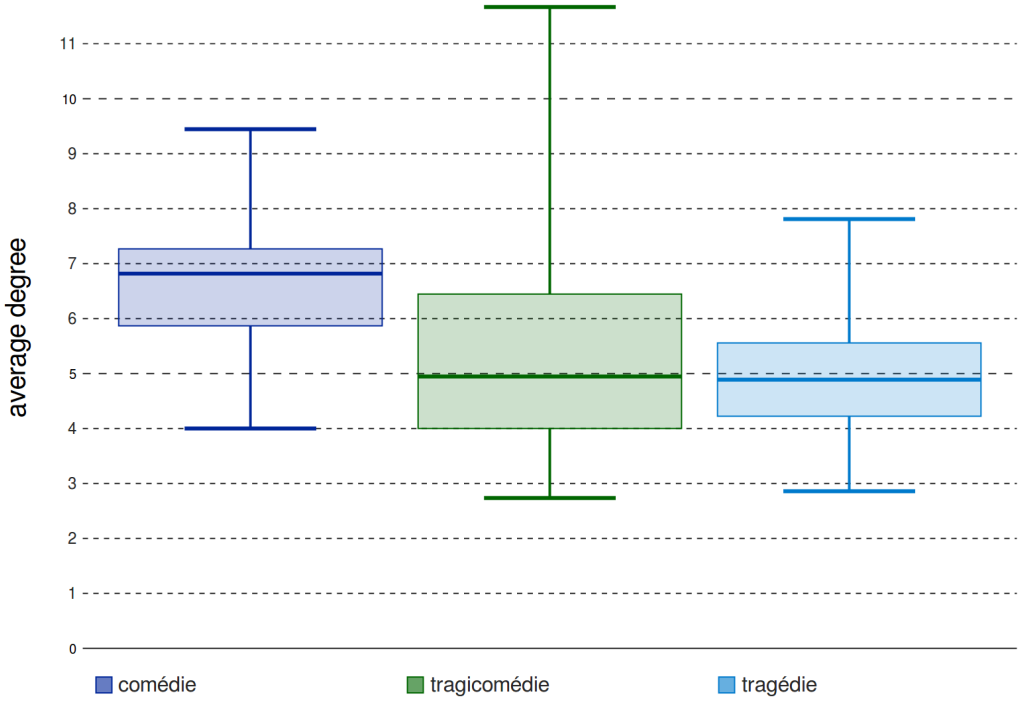
44 This kind of information can be used in several ways. First of all, a network of interactions can be constructed based on how closely related the different characters in a play are. One way of operationalizing this notion of "relatedness" is to consider how many words a given character speaks to the other characters present in the same scene. A file format derived from the original TEI format, the so-called "Zwischenformat" (see Kampkaspar, Fischer, and Trilcke 2015), which we also make available for the collection of plays, makes this type of analysis particularly easy. Figure 2 shows the weighted network for Jean Racine's tragedy *Britannicus* (1669).

Figure 2. Character network based on number of words spoken in mutual presence (represented by the thickness of lines), for Jean Racine's tragedy *Britannicus* (1669).



45 This graph clearly shows, for example, that Britannicus and Néron interact surprisingly little despite being direct opponents. Rather, their conflict, which concerns Junie, also passes via Junie. Also notable is the fact that Agrippine interacts more intensely with Burrhus, the tutor of Néron, than directly with Néron. This type of analysis becomes more interesting, however, when looking not at individual networks, but at trends and patterns in key network indicators (such as network density or average degree; see, e.g., Newman 2003) across a larger collection of plays.

Figure 3. Box plots for average degree in the character networks, for three subgenres.



46 Figure 3 shows box plots of average degree for the complete collection of 100 French plays separated by subgenre. Average degree is a measure of how many edges there are between nodes in a network. In the case of our analysis of plays, this means that an average degree indicates the number of other characters each character interacts with, on average, in a given play. The results show that in comedies, characters appear to have more connections with each other, on average, than in both tragedies and tragicomedies. Also, the median of values is similar for tragedy and tragicomedy, but the range of values is much higher in tragicomedy than in the other two subgenres, indicating a wider range of structural choices in this subgenre.

5.3 Textometric Analysis

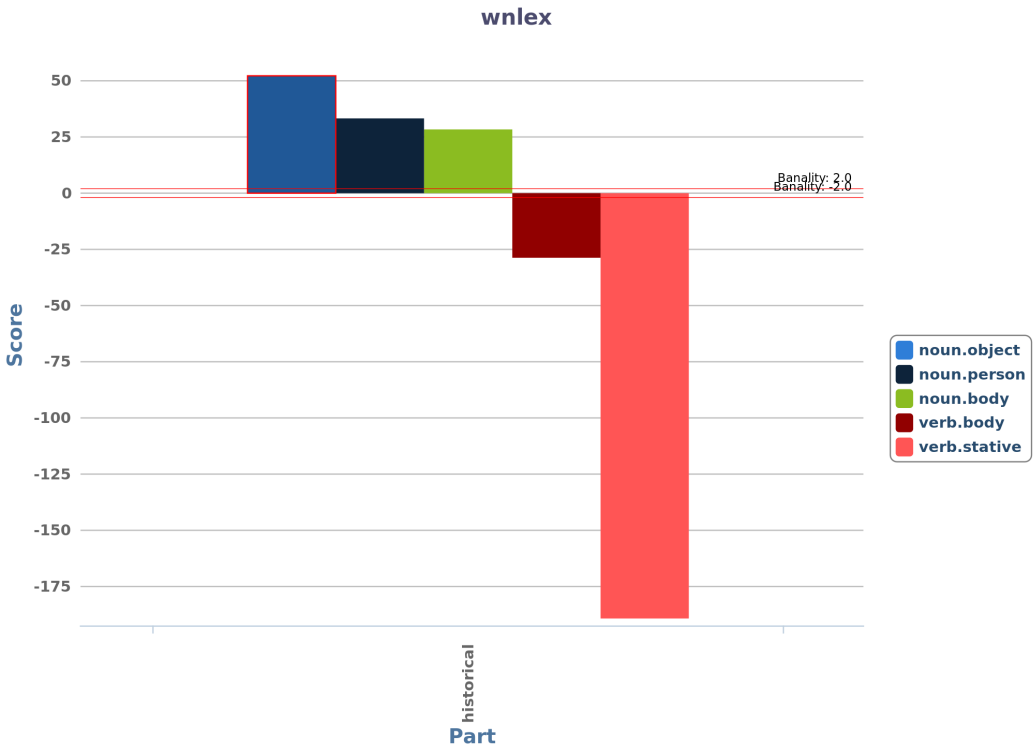
47 Yet another way to use the text collections is to conduct stylistic analyses with TXM, a tool developed in France starting in 2007 in the framework of an approach called *textométrie* (see Heiden 2010). TXM is freely available[12] and supports the analysis of large text corpora on lexical

and morphological levels, also taking metadata into consideration. Here, the XML format with annotations from FreeLing and WordNet is used instead of the built-in TreeTagger support of TXM (see section 2.2).

48 With the help of TXM, a corpus can be established by importing single files in one of the possible formats. The metadata may include any type of information about the texts, such as title, author, year and country of publication, or literary subgenre. This information can then be used in the analyses, for example when dividing a corpus into a partition based on certain metadata values (e.g., different subgenres). One type of analysis that TXM supports is called *specificities*. Similarly to contrastive analyses using a t-test or rank-sum test (see Lijffijt et al. 2016) or measures such as Zeta (see Schöch 2018), this analysis determines forms that are distinctive for one part in comparison to other parts of a text collection, for example which word forms are specific to the subgenre "historical novel" compared to other novelistic subgenres in a text collection partitioned by subgenre.

49 Figure 4 shows an example of a specificities analysis with TXM where the collection of 24 Spanish-American novels has been partitioned by subgenre. The distinctive features of historical novels were calculated in comparison to the novels of other subgenres, using WordNet semantic classes as features (so-called *lexnames*, for lexicographer file names[13]). The five most distinctive features are shown in the figure. Nouns denoting natural objects (`noun.object`), people (`noun.people`), and body parts (`noun.body`) have particularly high values for historical novels. Verbs of grooming, dressing, and bodily care (`verb.body`) and verbs of being, having, and spatial relations (`verb.stative`) are underrepresented in historical novels when compared to novels from other subgenres. Interestingly, by far the most distinctive feature (`verb.stative`) is one that is particularly weak in historical novels.

Figure 4. Specificities analysis with TXM.

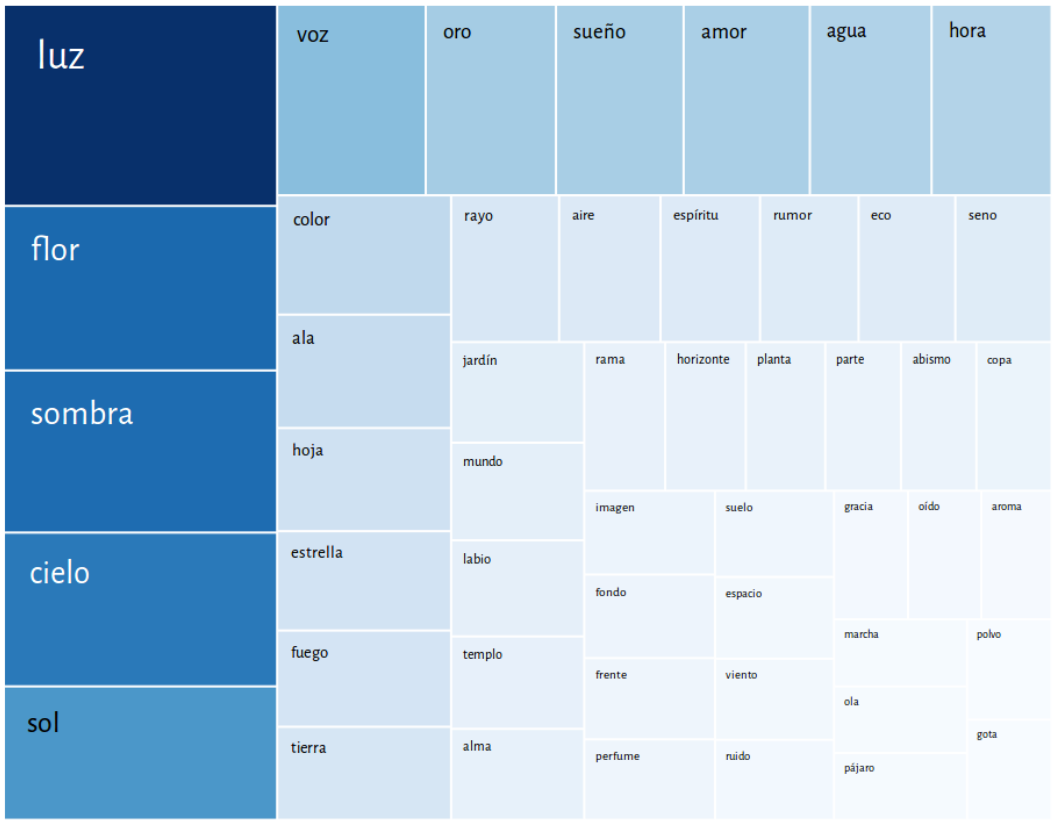


5.4 Topic Modeling

50 Finally, the collections support analyses using topic modeling. Owing to the limited size of the collections, analytical work based on topic modeling should be regarded as proof-of-concept analyses rather than permitting substantial insight into topic-based patterns and evolutions in French, Spanish, Spanish-American, or Portuguese literature. The basic procedure (as supported, for example, by the Topic Modeling Workflow set of Python scripts made available by the CLiGS group[14]) consists of four or five steps: preparing the texts, performing the actual topic modeling, optionally evaluating the model before proceeding, postprocessing the raw output data, and visualizing the results. Text preparation primarily involves linguistic annotation (such as providing information on lemmas and parts of speech), selecting a subset of lemmatized word forms based on part-of-speech information and/or filtering out stop words (such as high-frequency function words, words appearing only once in the collection, and named entities), and splitting the rather

long novels into shorter segments. Topic modeling itself is then performed using, for instance, MALLET to obtain a set of probability distributions, notably the probabilities of words in topics and the probabilities of topics in text segments.

Figure 5. Treemap visualization of a single topic with its top 50 words (from a 100-topic model of 192 Spanish-American novels). Box size and color reflect the probability of the word in the topic: the more probable a word, the larger and darker the box.



51 Postprocessing primarily connects the raw output from MALLET to metadata about each segment (regarding the novel each segment belongs to and hence information such as title, author, year of publication, or subgenre). Based on these data, visualizations can then be generated to show the topics themselves (for instance, as word clouds or treemaps, as illustrated in figure 5) and their distributional patterns in the collection (for instance, using heatmaps for topic distributions over subgenres, bar charts to show top topics for a given novel or author, or line plots to show the

evolution over time of one or several topics). For an example of such analyses using a collection of French plays, see Schöch 2017, and for another example using Spanish and Spanish-American novels, see Schöch et al. 2016.

6. Conclusion

52 In this paper, we hope to have documented how we compiled, annotated, and published the collections of literary texts included in the CLiGS textbox, to have provided a rationale for why we proceeded in this way, and to have shown several ways in which the text collections can be used for research in literary studies. In conclusion, we offer a few thoughts emerging from our activities in collection-building for digital research in Romance studies.

53 When building and using the textbox collections, agreeing on common formats and procedures of text preparation and encoding has been crucial. For the CLiGS group, it has been helpful to agree as much as possible on a common strategy—and a realizable one—in order to share and bundle experiences and efforts. Also, existing best practices, such as recommended subsets of the TEI and existing infrastructure components like GitHub and Zenodo, were essential in establishing the setup for the textbox as described in this article. Another advantage of a common strategy is that investigations into a given research question can make use of several of the different collections at a time. The research group has already benefited from this when combining texts from Spain and Spanish-America for textual analyses, for example in the study using topic modeling for genre analysis mentioned above (see Schöch et al. 2016). Also, code developed to create, transform, and analyze textual data can be reused across all collections, which has allowed the development of the CLiGS toolbox.[15] This is the tool-oriented counterpart of the textbox: a collection of Python scripts covering various aspects of text curation, collection building, and simple analyses.

54 The current state of access to literary texts in digital format for researchers in Romance studies appears to be far from ideal. Although many texts are in principle available in an electronic format, there are a number of caveats. Often, no full text is offered, or the quality of the full text is not very good. In many cases, texts are offered as image-based PDF files. In some cases, e-books are only presented in proprietary formats (e.g., Mobipocket or Kindle). Sometimes there are access restrictions, for example based on institutional affiliations or the country of residence, even when the texts are in the public domain. More generally, the landscape is highly fragmented;

researchers hoping to build substantial collections need to rely on multiple, heterogeneous sources. Researchers working with literary texts in Romance languages and external to the group have already shown interest in reusing the CLiGS collections. This shows that access to a large number of historical literary texts of a certain language, region, and period which have been prepared so as to be suitable for quantitative text analyses is desirable and cannot yet be taken for granted.

55 There are at least four major challenges in providing access to research data: standardization, openness, sustainability, and discoverability. An important strategy to help mitigate the adverse effects of the fragmented landscape of available texts is the use of standardized formats for text preparation and encoding, along with maximal openness in terms of technical convenience and of licenses when publishing data and metadata. Using well-supported research data repositories to archive research data should ensure the long-term availability of the data. We have pointed out possible solutions with our text collections, and we hope to encourage others who have prepared electronic versions of literary texts for research (or plan to do so) to share their collections in a similar manner, to make the material available to a wide audience at an interdisciplinary and international scale. It is an even greater challenge, however, to improve the discoverability of text collections in a field such as Romance studies rooted in several continents and numerous countries. Currently, there does not appear to be a way to ensure that smaller text collections like the ones presented here become and remain findable and visible inside and beyond the community of Romance studies, despite efforts such as DARIAH-DE's Collection Registry[16] or the activities of the Research Data Alliance (RDA).[17] Finally, a recently launched effort aimed at German-speaking scholars in Romance studies is the "Resources" registry for datasets and tools that forms part of the information platform romanistik.de.[18]

APPENDIXES

Appendix 1. Sources for Literary Texts in Romance Languages[19]

The following is a list of sources for digital full-text versions of literary texts in several Romance languages, notably French and Spanish, but also Italian and Portuguese. Note that this is a non-exhaustive list containing only those repositories and digital libraries which provide freely available digital full-text versions of literary texts (not image-based PDF versions as many digital libraries do) by more than one author.

Source name	URL	Languages	Formats
Project Gutenberg	https://www.gutenberg.org/	French, Spanish, Italian, Portuguese, among others	HTML, EPUB, PDF, plain text
Wikisource	https://wikisource.org	French, Spanish, Italian, Portuguese, among others	HTML
ManyBooks	http://manybooks.net	French, Spanish, Italian, Portuguese, among others	HTML, EPUB, RTF, PDF, among others
Domínio Público	http://www.dominiopublico.gov.br	Portuguese, Spanish, French, Italian, among others	HTML, PDF

virtualbooks	http://www.virtualbooks.com.br	Portuguese, Spanish, French, Italian, among others	PDF (text)
ATHENA	http://athena.unige.ch/athena/	French	HTML, RTF
Bibliothèque ABU	http://abu.cnam.fr/	French	HTML, plain text
Bibliothèque dramatique du CELLF	http://bibdramatique.paris-sorbonne.fr/	French	EPUB, HTML, PDF, TEI XML, plain text
Ebooks libres et gratuits	http://www.ebooksgratuits.com/	French	HTML, EPUB, PDF, among others
ÉFÉLÉ	http://efele.net/ebooks/	French	EPUB
La Bibliothèque électronique du Québec	http://beq.ebooksgratuits.com/	French	EPUB, PDF, among others
Théâtre classique	http://www.theatre-classique.fr	French	HTML, PDF, plain text (TEI XML P4)
Biblioteca Virtual Antorcha	http://www.antorcha.net/index/biblioteca.html	Spanish	HTML
Biblioteca Virtual Miguel de Cervantes	http://www.cervantesvirtual.com/	Spanish	HTML, PDF

Comedias	http://www.wordpress.comedias.org	Spanish	HTML, PDF
Corpus of Spanish Golden-Age Sonnets	https://github.com/bncolorado/CorpusSonetosSigloDeOro	Spanish	TEI XML
epublibre	https://www.epublibre.org	Spanish	EPUB
El Libro Total. La Biblioteca de América	http://www.ellibrototal.com	Spanish	HTML
IMPACT-es diachronic corpus	http://www.digitisation.eu/tools-resources/language-resources/impact-es/	Spanish	TEI XML
Biblioteca italiana	http://www.bibliotecaitaliana.it/	Italian	XML
Liber Liber	http://www.liberliber.it/	Italian	EPUB, ODT, PDF
Biblioteca Digital Camões	http://cvc.instituto-camoes.pt/conhecer/biblioteca-digital-camoes.html	Portuguese	PDF (text)
LELIVROS	http://lelivros.love	Portuguese	EPUB, PDF, among others
Luso Livros	https://www.luso-livros.net/	Portuguese	EPUB, PDF

BIBLIOGRAPHY

Burnard, Lou. 2014. “Introduction” to *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources.* Encyclopédie Numérique 3. Marseille: OpenEdition Press. http://books.openedition.org/oep/679. doi:10.4000/books.oep.679.

Burnard, Lou, and C. M. Sperberg-McQueen. 2012. "TEI Lite: Encoding for Interchange: An Introduction to the TEI." Final Revised Edition for TEI P5. http://www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_lite.doc.html.

Calvo Tello, José. 2016. "Entendiendo Delta desde las Humanidades." *Caracteres. Estudios culturales y críticos de la esfera digital* 5, no. 1: 140–76. http://revistacaracteres.net/revista/vol5n1mayo2016/entendiendo-delta/.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 8, no. 1: 107–21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. "Understanding and Explaining Delta Measures for Authorship Attribution." *Digital Scholarship in the Humanities* 32, issue suppl_2, ii4–ii16. doi:10.1093/llc/fqx023.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, MA: MIT Press.

Fièvre, Paul, ed. 2007–2018. "Théâtre classique." http://www.theatre-classique.fr.

Haaf, Susanne, Alexander Geyken, and Frank Wiegand. 2014/15. "The DTA 'Base Format': A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources." *Journal of the Text Encoding Initiative* 8. https://journals.openedition.org/jtei/1114; doi:10.4000/jtei.1114.

Heiden, Serge. 2010. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation - PACLIC24*, edited by Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, and Yasunari Harada, 389–98. Sendai: Waseda University. https://www.aclweb.org/anthology/Y10-1044. Also available at https://halshs.archives-ouvertes.fr/halshs-00549764/en.

Henny, Ulrike, and Christof Schöch. 2016. "How Good Are Our Texts, Really? Quality Assurance for Literary Texts from Various Sources." Scholarly blog post. February 27. Würzburg: CLiGS. http://cligs.hypotheses.org/371.

Kampkaspar, Dario, Frank Fischer, and Peer Trilcke. 2015. "Introducing Our 'Zwischenformat.'" *Network Analysis of Dramatic Texts*. Scholarly blog post. June 21. https://dlina.github.io/Introducing-Our-Zwischenformat/.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. 2016. "Significance Testing of Word Frequencies in Corpora." *Digital Scholarship in the Humanities* 31, no. 2: 374–97. doi:10.1093/llc/fqu064.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38, no. 11: 39–41. doi:10.1145/219717.219748.

Mueller, Martin, Sebastian Rahtz, Bryan Pytlik Zillig, James Cummings, and Magdalena Turska. 2015. "TEI Simple: An Introduction." Version 0.9, August. http://htmlpreview.github.io/?https://github.com/TEIC/TEI-Simple/blob/master/teisimple.html.

National Information Standards Organization (NISO). 2004. *Understanding Metadata*. Bethesda, MD: NISO. https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf.

Newman, M. E. J. 2003. "The Structure and Function of Complex Networks." *SIAM Review* 45, no. 2: 167–256. doi:10.1137/S003614450342480.

Nielsen, Lars Holm. 2013. "ZENODO - An Innovative Service for Sharing All Research Outputs." Presented at Joint OpenAIRE/LIBER Workshop, Ghent, Belgium, May 28. doi:10.5281/zenodo.6815.

Padró, Lluís, and Evgeny Stanislovsky. 2012. "FreeLing 3.0: Towards Wider Multilinguality." In *Proceedings of LREC 2012, Eighth International Conference on Language Resources and Evaluation*. N.p.: European Language Resources Association. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf.

Schöch, Christof. 2017. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11, no. 2. http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

———. 2018. "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie." In *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*, edited by Toni Bernhart, Marcus Willand, Sandra Richter, and Andrea Albrecht, 77–94. Berlin: de Gruyter. doi: 10.1515/9783110523300-004.

Schöch, Christof, Ulrike Henny, José Calvo, Daniel Schlör, and Stefanie Popp. 2016. "Topic, Genre, Text. Topics im Textverlauf von Untergattungen des Spanischen und Hispanoamerikanischen Romans (1880–1930)," *Konferenzabstracts, DHd 2016. Modellierung—Vernetzung—Visualisierung: Die Digital Humanities als fächerübergreifendes Forschungsparadigma*, 235–39. Leipzig: nisaba verlag. http://dhd2016.de/boa.pdf.

Smith, Peter W. H., and W. Aldridge. 2011. "Improving Authorship Attribution: Optimizing Burrows' Delta Method." *Journal of Quantitative Linguistics* 18 (1): 63–88. doi:10.1080/09296174.2011.533591.

TEI Consortium. 2018. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.4.0. Last updated July 23, 2018. N.p.: TEI Consortium. https://tei-c.org/Vault/P5/3.4.0/doc/tei-p5-doc/en/html/.

NOTES

1 CLiGS web page in German, accessed May 11, 2018, http://www.cligs.hypotheses.org; in English, accessed July 10, 2019, https://cligs.hypotheses.org/sprachen/english.

2 Christof Schöch, Ulrike, José Calvo, and Katrin Betz, "cligs/textbox: Almost Summer Release," May 28, 2018, doi:10.5281/zenodo.1254483.

3 CLiGS Reference Repository, latest commit May 8, 2018, http://github.com/cligs/reference/.

4 This way, the encoded files are more easily comparable and treatable as a text collection for information extraction and analysis. When fewer kinds of elements are used, path expressions can be simpler. To extract all block-level elements inside text divisions, for example—a common task in text analysis—not all of the possible TEI elements have to be considered in a query.

5 Last modified March 2, 2018, http://textometrie.ens-lyon.fr/.

6 "Network Analysis of Dramatic Texts," Digital Literary Network Analysis group, last modified March 6, 2018, https://dlina.github.io/.

7 Last modified May 27, 2018, http://viaf.org/.

8 Note that no full record of changes is required here because the collections are managed on GitHub with fully automatic version control. The subsequent releases are archived on Zenodo and can be compared to each other.

9 TEI Consortium 2018, sec. 2.4.3, "The Text Classification," https://tei-c.org/Vault/P5/3.4.0/doc/tei-p5-doc/en/html/HD.html#HD43.

10 Accessed July 10, 2019, https://www.base-search.net/.

11 It may be noted that publishing an article such as the present one is also motivated, in part, by the desire to provide an opportunity to others to give credit to the editors of the text collections.

12 See "Textométrie welcome page," last updated July 5, 2019, http://textometrie.ens-lyon.fr/?lang=fr.

13 "lexnames(5WN)," *WordNet: A Lexical Database for English*, accessed June 28, 2019, https://wordnet.princeton.edu/documentation/lexnames5wn.

14 See "Topic Modeling Workflow in Python," CLiGS GitHub site, latest commit on May 8, 2017, http://github.com/cligs/tmw; Christof Schöch, "cligs/tmw: Topic Modeling Genre Release," version v0.3.0, Zenodo, April 3, 2017, doi:10.5281/zenodo.439975.

15 See cligs/toolbox, latest commit on September 17, 2018, http://github.com/cligs/toolbox.

16 Accessed May 30, 2018, http://colreg.de.dariah.eu/.

17 Accessed May 30, 2018, https://rd-alliance.org/.

18 See "Forschung: Ressourcen," accessed May 30, 2018, https://romanistik.de/res.

19 Not all the websites mentioned below are stable. Some of the URLs are prone to disappear or to change in the course of time.

AUTHORS

CHRISTOF SCHÖCH

Christof Schöch is professor of Digital Humanities and co-director of the Trier Center for Digital Humanities at the University of Trier, Germany. He has led the early-career research group Computational Literary Genre Stylistics (CLiGS) at the Department for Literary Computing, University of Würzburg, Germany. His interests in research and teaching are located at the confluence of French literary studies and digital humanities, especially digital editing and quantitative text analysis. Website: http://www.christof-schoech.de/en.

JOSÉ CALVO TELLO

José Calvo Tello is a PhD candidate at the Department for Literary Computing, University of Würzburg, Germany, and a member of the early-career research group Computational Literary Genre Stylistics (CLiGS). He studied Spanish literature and linguistics and has since been involved in projects that apply new technologies to the study of texts in Spanish. He has been involved in the collection and publication of several corpora and collections of texts of Spanish Literature. Currently he is analyzing the subgenre of the novels of the Spanish *Edad de Plata* (Silver Age), modifying and applying stylometric methods. His website is http://www.morethanbooks.eu/.

ULRIKE HENNY-KRAHMER

Ulrike Henny-Krahmer is a PhD candidate at the Department for Literary Computing, University of Würzburg, Germany, and a member of the early-career research group Computational Literary Genre Stylistics (CLiGS). Her current work is about subgenres of the nineteenth-century Spanish American novel, with a focus on the analysis of topics in relation to other, stylistic means. She is further interested in digital scholarly editions and text collections and the evaluation of digital scholarship. She is a member of the Institute for Documentology and Scholarly Editing (IDE) and a co-editor of the review journal *ride*. Website: http://ulrike-henny.de.

STEFANIE POPP

Stefanie Popp was a member of the early-career research group Computational Literary Genre Stylistics (CLiGS) from April 2015 to April 2016. During this time, her research focussed on French literature and, in particular, on fantastic novels and novellas of the nineteenth century.